

# (DE)<sup>2</sup>CO: Deep Depth Colorization

F. M. Carlucci\*, P. Russo\*, S. M. Baharlou, and B. Caputo<sup>1</sup>

**Abstract**—Object recognition on depth images using convolutional neural networks requires mapping the data collected with depth sensors into three dimensional channels. This makes them processable by deep architectures, pre-trained over large scale RGB databases like ImageNet. Current mappings are based on heuristic assumptions over what depth properties should be most preserved, resulting often in cumbersome data visualizations, and likely in sub-optimal recognition results. Here we take an alternative route and we attempt instead to *learn* an optimal colorization mapping for any given pre-trained architecture, using as training data a reference RGB-D database. We propose a deep network architecture, exploiting the residual paradigm, that learns how to map depth data to three channel images from a reference database. A qualitative analysis of the images obtained with this approach clearly indicates that learning the optimal mapping for depth data preserves the richness of depth information much better than hand-crafted approaches currently in use. Experiments on the Washington, JHUIT-50 and BigBIRD public benchmark databases, using AlexNet, VGG-16, GoogleNet, ResNet and SqueezeNet, clearly showcase the power of our approach, with gains in performance of up to 17% compared to the state of the art.

## I. INTRODUCTION

Robots need to recognize what they see around them to be able to act and interact with it. Recognition needs to be carried out in the RGB domain, that mostly captures the visual appearance of things related to their reflectance properties, as well as in the depth domain, which provides information about the shape and silhouette of objects, supporting at the same time recognition and interaction with items. The current mainstream state of the art approaches for object recognition are based on Convolutional Neural Networks (CNNs, [1]), which use end-to-end architectures achieving feature learning and classification at the same time. Some notable advantages of these networks are their ability to reach much higher accuracies on basically any visual recognition problem, compared to what would be achievable with heuristic methods; their being domain-independent, and their conceptual simplicity. Despite these advantages, they also present some limitations, such as high computational cost, long training time and the demand for large datasets, among others.

This last issue has so far proved crucial in the attempts to leverage over the spectacular success of CNNs over RGB-

based object categorization [2], [3] in the depth domain. Being CNNs data-hungry algorithms, the availability of very large scale annotated data collections is crucial for their success, and architectures trained over ImageNet [4] are the cornerstone of the vast majority of CNN-based recognition methods. Besides the notable exception of [5], the mainstream approach for using CNNs on depth-based object classification has been to design a mapping able to make the input channel compatible with the data distribution expected by the chosen architecture. Specifically, the depth pixel represents the distance between the sensor and the object, and the entire depth map can be represented by a grayscale image with a float type. On the other hand, CNNs architectures are trained on RGB images with three integer channels. Following recent efforts in transfer learning [6], [7], [8] that made it possible to use depth data with CNN pre-trained on a database of a different modality, several authors proposed hand-crafted mappings to colorize depth data, obtaining impressive improvements in classification over shallow features [9], [10].

We argue that this strategy is sub-optimal, as it does not truly exploit the power of end-to-end convolutional networks. Inspired by recent work on colorization of grayscale photographs [11], [12], [13], we propose a deep depth colorization architecture able to learn the optimal colorization mapping for any given pre-trained architecture. Our deep colorization network takes advantage of the residual approach [14], learning how to map from depth to RGB by leveraging over a reference database (Figure 1, top), for a given architecture. After this training stage, the colorization network can be added on top of its reference pre-trained architecture, for any object classification task (Figure 1, bottom). We call our network (DE)<sup>2</sup>CO: DEep DEpth Colorization. A qualitative analysis of the colorized depth images obtained with (DE)<sup>2</sup>CO clearly shows the superiority of our method, compared with the very popular ColorJet mapping [10]. Experiments performed over three different databases and five different kind of architectures show the clear power of (DE)<sup>2</sup>CO, with increases in performance of up to 17% compared to ColorJet, that, given its popularity in the literature and the benchmark results reported in [10], can be considered as the off-the-shelf state of the art depth colorization mapping. Upon acceptance of the paper, we will make available to the community the (DE)<sup>2</sup>CO modules, for the different architectures employed in this paper.

The rest of the paper is organized as follows: after a review of relevant previous work (section II), section III describes our colorization network. Section IV reports our experimental findings, and we conclude the paper with a

This work was partially supported by the ERC grant 637076 - RoboExNovo (B.C.), and the CHIST-ERA project ALOOF (B.C, F. M. C., P. R.).

\* Equal contribution

<sup>1</sup>All authors are at the Department of Computer, Control, and Management Engineering Antonio Ruberti at Sapienza University of Rome, Rome, Italy { fabiom.carlucci,baharlou,prusso,caputo } @dis.uniroma1.it

summary and a discussion on future research avenues.

## II. RELATED WORK

Since the spectacular success of Krizhevsky’s AlexNet [3], CNNs have become the dominant learning paradigm in visual object recognition. Several architectures have been proposed over the last years, each bringing new flavors to the community. Simonyan and Zisserman [15] investigated the effect of increasing the network depth, using an architecture with small ( $3 \times 3$ ) filter maps in each layer and combining them as a sequence of convolutions. GoogLeNet [2] also increased the depth and width of the network while restraining the computational budget, with a dramatic reduction in the number of parameters of the architecture. He et al. [14] proposed a residual learning approach using a batch normalization layer and special skip connections for training deeper architectures, showing an impressive success in the 2015 ImageNet visual recognition challenge. Iandola et al. [16] proposed a compact architecture able to achieve AlexNet’s performance with fifty times fewer parameters, with a considerable computational advantage. All these architectures will be used in this work, to assess the generality of our algorithm.

Lately, several authors attempted to take advantage of pre-trained CNNs to perform RGB-D detection and recognition. Gupta et al [17] applied the R-CNN approach [18] for object detection to depth images; Schwarz et al [19] used two concurrent CNNs pre-trained on ImageNet to extract features from RGB-D instances. The extracted features were then followed by an SVM classifier and regressor for category prediction and pose estimation. A work analogous to [19] was proposed by Eitel et al [10] where they used a multi-modal CNN architecture for RGB-D object recognition. All these works, and many others [20], [5], make use of an ad-hoc mapping for converting depth images into three channels. This conversion is vital as the dataset has to be compatible with the pre-trained CNN. Depth data is encoded as a 2D array where each element represents an approximate distance between the sensor and the object. Depth information is often depicted and stored as a single monochrome image. Compared to regular RGB cameras, the depth resolution is relatively low, especially when the frame is cropped to focus on a particular object.

The simplest method for converting an 11-bit depth image to a 24-bit RGB image is to decrease the accuracy from 11-bit to 8-bit, and copy the resulting image to each RGB channel. Alternatively, a more robust way proposed by [10] is to normalize the depth data between zero and one, then extend it into three channels. Zero values indicate the regions where the depth information is not available; hence, they can be excluded from the normalization stage for more accurate results. As suggested by Bo et al [21], another approach is to re-interpret the surface normals as the RGB values, a method similar to the normal-mapping technique widely used in the computer graphic community. This method does not deal with the intensity of the pixels, instead, the relation of a pixel w.r.t its neighbors is considered.

Schwarz et al [19] proposed a colorization pipeline where colors are assigned to the image pixels according to the distance of the vertexes of a rendered mesh to the center of the object. HHA, proposed by Gupta et al [17], is a mapping where one channel encodes the horizontal disparity, one the height above ground and the third the pixelwise angle between the surface normal and the gravity vector. Besides the naive grayscale method, the rest of the mentioned colorization schemes are computationally expensive. Eitel et al [10] used a simple color mapping technique known as *ColorJet*, showing that this simple method outperformed more sophisticated approaches. In the rest of this work, we will consider ColorJet as the reference heuristic colorization method against which to compare.

Our work is also related to the colorization of grayscale images using deep nets. Cheng et al [13] proposed a colorization pipeline based on three different hand-designed feature extractors to determine the features from different levels of an input image. Larsson et al [12] used an architecture consisting of two parts. The first part is a fully convolutional version of VGG-16 used as feature extractor, and the second part is a fully-connected layer with 1024 channels predicting the distributions of hue and the chroma for each pixel given its feature descriptors from the previous level. Iizuka et al [11] proposed an end-to-end network able to learn global and local features, exploiting the classification labels for better image colorization. Their architecture consists of several networks followed by fusion layer for the colorization task. Our work differs from this last research thread in the specific architecture proposed, and in its main goal, as here we are interested in learning optimal mapping for categorization rather than for colorization of grayscale images.

## III. COLORIZATION OF DEPTH IMAGES

Although depth and RGB are modalities with significant differences, they also share enough similarities (edges, gradients, shapes) to make it plausible that convolutional filters learned from RGB data could be re-used effectively for representing colorized depth images. The approach currently adopted in the literature consists of designing ad-hoc colorization algorithms, as revised in the previous section. We refer to these kind of approaches as *shallow depth colorization* in the rest of the paper. In 2015, Eitel et al [10] proposed ColorJet, an effective yet simple shallow depth colorization method combining state of the art performances on the Washington database with a very low computational cost. Since then, ColorJet has become the off-the-shelf state of the art shallow colorization approach for depth-based object recognition [20], [5].

In the rest of this section we first describe the ColorJet algorithm (section III-A), then we describe our deep approach to depth colorization (section III-B). To the best of our knowledge, (DE)<sup>2</sup>CO is the first deep colorization architecture applied successfully to depth images.

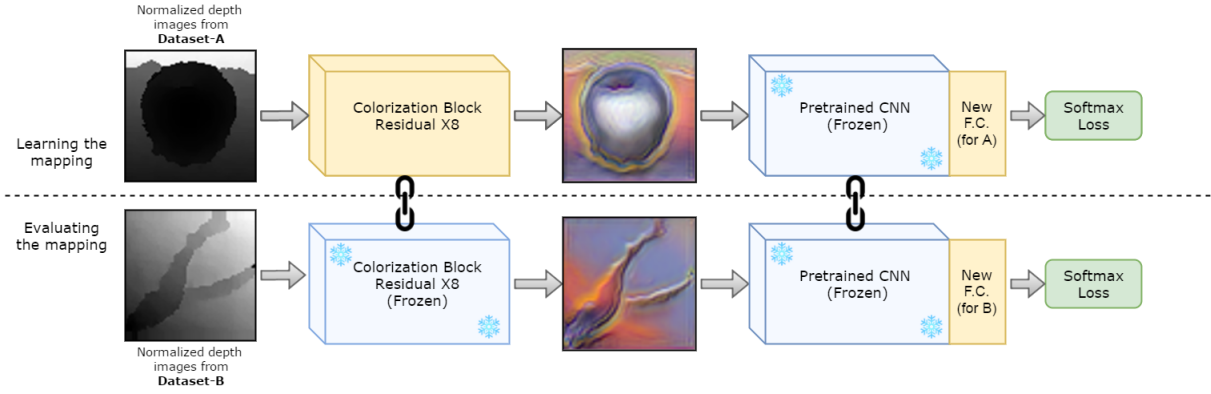


Fig. 1: The (DE)<sup>2</sup>CO pipeline is composed by two phases. First, we learn the mapping, from depth to color, which maximizes the discrimination capabilities of a network pre trained on ImageNet. During this step the network is frozen and we are only learning the mapping and the final layer (which must be relearned, as the classes are different). We then evaluate the colorization on a **different** depth dataset: here we also freeze the colorization network and only train a new final layer for the testbed dataset.

#### A. Shallow Depth Colorization: ColorJet

The ColorJet technique works by assigning different colors to different depth values. The original depth map is firstly normalized between 0-255 values. Then the colorization works by mapping the lowest value to the blue channel and the highest value to the red channel. The value in the middle is mapped to green and the intermediate values are arranged accordingly [10]. The resulting image exploits the full RGB spectrum, with the intent of leveraging at best the filters learned by deep networks trained on very large scale RGB datasets like ImageNet. Although simple, the approach gave very strong results when tested on the Washington database, and when deployed on a robot platform. Still, ColorJet was not designed to create realistic looking RGB images for the objects depicted in the original depth data (Figure 3, bottom row). This raises the question whether this shallow mapping, even though more effective than other methods presented in the literature, might be sub-optimal. In the next section we will show that by fully embracing the end-to-end philosophy at the core of deep learning, it is indeed possible to achieve significantly higher recognition performances while at the same time producing more realistic colorized images.

#### B. Deep Depth Colorization: (DE)<sup>2</sup>CO

(DE)<sup>2</sup>CO, our deep depth colorization method, consists of feeding the depth maps, normalized into grayscale images, to a *colorization network* linked to a standard CNN architecture, pre-trained on ImageNet.

Given the success of deep colorization networks from grayscale images, we first tested existing architectures in this context [22]. Extensive tests showed that while the visual appearance of the colorized images was very good, the recognition performances obtained when combining such network with pre-trained RGB architectures was not competitive with shallow colorization methods. Inspired by the generator network proposed in [23], we propose here instead a *residual* convolutional architecture (Figure 2).

Our architecture works as follows: the 1x228x228 input depth map, reduced to 64x57x57 size by a conv&pool layer, passes through a sequence of 8 residual blocks, composed by 2 small convolutions with batch normalization layers and leakyRelu as non linearities. The last residual block output is convolved by a three features convolution to form the 3 channels image output. Its resolution is brought back to 228x228 by a *deconvolution* (upsampling) layer.

Our whole system for object recognition in the depth domain using deep networks pre-trained over RGB images can be summarized as follows: the entire network, composed by (DE)<sup>2</sup>CO and the classification network of choice, is trained on an annotated reference depth image dataset. The weights of the chosen classification network are kept frozen in their pre-trained state, as the only layer that needs to be retrained is the last fully connected layer connected to the softmax layer. Meanwhile, the weights of (DE)<sup>2</sup>CO are updated until convergence.

After this step, the depth colorization network has learned the mapping that maximizes the classification accuracy on the reference training dataset. It can now be used to colorize *any* depth image, from any data collection. Figure 3, top rows, shows exemplar images colorized with (DE)<sup>2</sup>CO trained over different reference databases, in combination with two different architectures (CaffeNet, an implementation variant of AlexNet, and VGG-16 [15]). We see that, compared to the images obtained with ColorJet, our results are more nuanced and expressive, especially when the training is conducted on databases containing varied and perceptually rich visual samples as is the case for the Washington database. In the next section we will show how this qualitative advantage translates also into a numerical advantage, i.e. how learning (DE)<sup>2</sup>CO on one dataset and performing depth-based object recognition on another leads to a very significant increase in performance compared to using ColorJet.

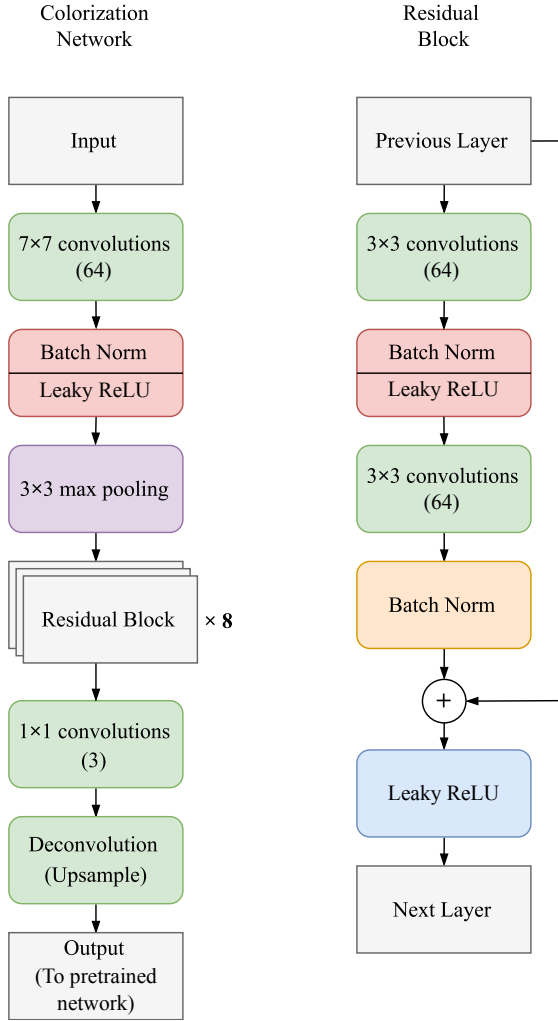


Fig. 2: Overview of the  $(DE)^2CO$  colorization network. On the left, we show the overall architecture; on the right, we show details of the residual block.

## IV. EXPERIMENTS

### A. Setup

We conducted experiments on the Washington RGB-D [24], the JHUIT-50 [25] and the BigBIRD [26] object datasets, which are the main public datasets for RGB-D object recognition. The first consists of 41,877 RGB-D images organized into 300 instances divided in 51 classes. Each object instance was positioned on a turntable and captured from three different viewpoints while rotating. Since two consecutive views are extremely similar, only 1 frame out of 5 is used for evaluation purposes. We performed experiments on the object categorization setting, where we followed the evaluation protocol defined in [24]. The JHUIT-50 is a challenging recent dataset that focuses on the problem of fine-grained classification. It contains 50 object instances, often very similar with each other (e.g. 9 different kinds of screwdrivers). As such, it presents different recognition challenges compared to the Washington database. Here we followed the evaluation procedure defined in [25].

BigBIRD is the biggest of the datasets we considered: it contains 121 object instances and 75.000 images. Unfortunately, it is an extremely unforgiving dataset for evaluating depth features: many objects are extremely similar, and many are boxes, which are indistinguishable without texture information. To partly mitigate this, we grouped together all classes annotated with the same first word: for example *nutrigrain apple cinnamon* and *nutrigrain blueberry* were grouped into *nutrigrain*. By applying this procedure, we reduced the number of classes to 61 (while keeping all of the samples). Another key issue is that objects are quite small (see figure 4) and because of this CNN based methods struggle greatly. For this reason, we used the object masks provided by [26] to crop around the object. Here we also follow the evaluation protocol defined in [25].

### B. Architectures

Our colorization scheme was evaluated on a total of five architectures. CaffeNet (a slight variant of the Alexnet [27]), VGG16 [15] and GoogleNet [2] were chosen because of their popularity within the robot vision community. We also tested the very recent ResNet50 [14] and the very efficient SqueezeNet v1.1 [16]; although they are not currently very used in the robotics domain, they present different advantages in terms of compactness and performance that promise to make them soon popular in this domain. In all cases we considered models pretrained on ImageNet [4], which we retrieved from Caffe’s *Model Zoo*<sup>1</sup>.

### C. Implementation details

1) *ColorJet*: To perform this colorization we normalized the data between 0 and 255 and then applied the ColorJet mapping using the OpenCV libraries<sup>2</sup>. Using this procedure we mapped red to the farthest away point, and blue to the closest. We also experimented with doing the opposite but did not observe any benefits. This baseline is evaluated by feeding the network the ColorJet images and retraining only the last layer.

2) *Learning  $(DE)^2CO$* : Training  $(DE)^2CO$  means minimizing the multinomial logistic loss of a network trained on RGB images. In practice, this means that our network is attached between the depth images and the pre-trained network, of which we freeze the weights of all but the last layer, which are relearned from scratch. We trained each network-dataset combination for 50 epochs using the Nesterov solver [28] and 0.007 starting learning rate (which is stepped down after 45%). During this phase, we used the whole *source* datasets, leaving aside only 10% of the samples for validation purposes.

3) *Assessing the colorization strategies*: Once  $(DE)^2CO$  has finished learning on dataset *A*, we tested its mapping on dataset *B* (which it had never seen before), by only training the new final layer (and freezing all the rest) for the classification task with softmax. Effectively we are using the pre-trained networks as feature extractors, as done in

<sup>1</sup><https://github.com/BVLC/caffe/wiki/Model-Zoo>

<sup>2</sup>“COLORMAPJET” from <http://opencv.org/>



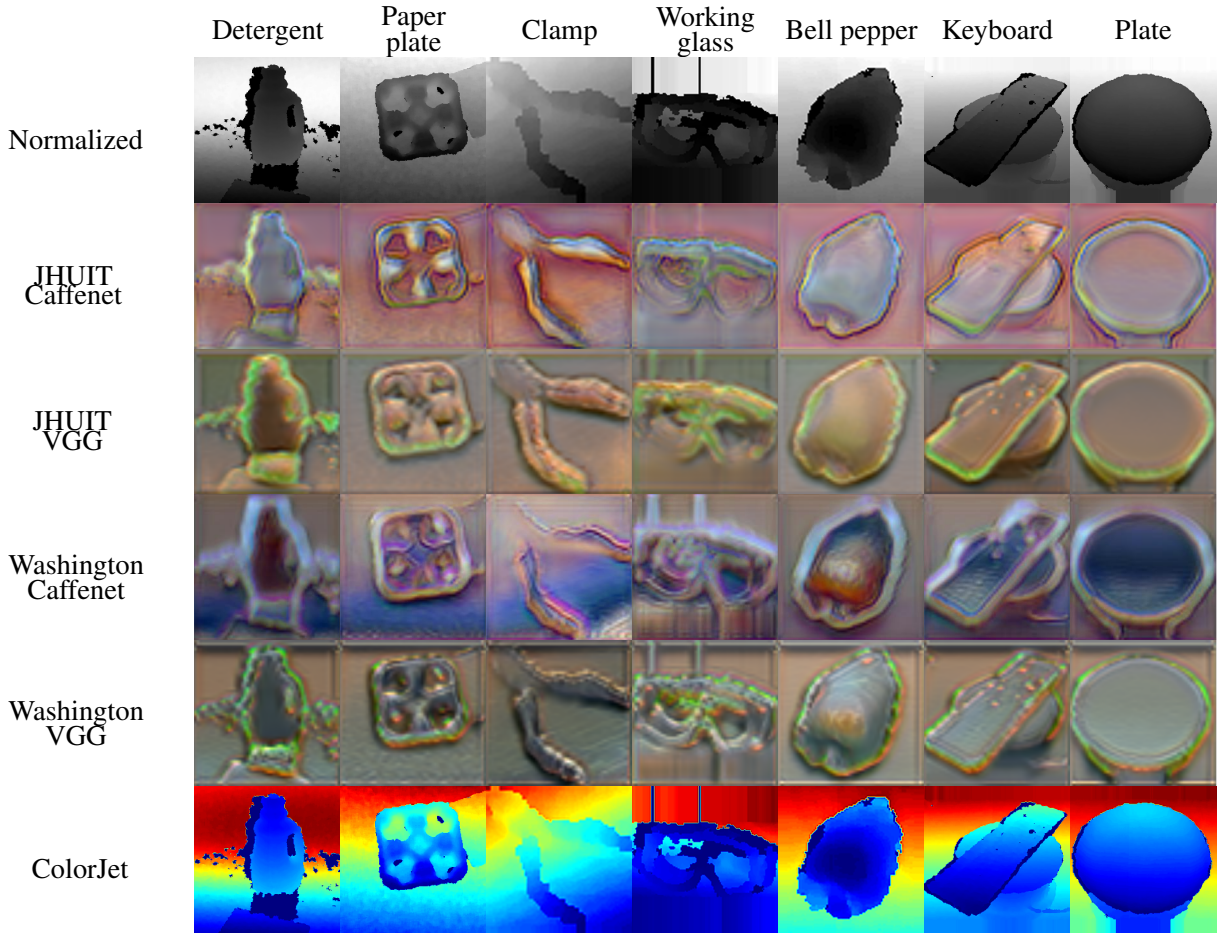


Fig. 3:  $(DE)^2CO$  colorizations applied on different objects taken from the Washington, JHUIT-50 and BigBIRD datasets. The top row represents the depth maps converted into 0-255 range grayscale image. From the second to the fourth row, we show the results of  $(DE)^2CO$  colorizations learned on different couples of training dataset/networks, applied on those depth maps. The bottom row shows the result of the ColorJet mapping. These images showcase  $(DE)^2CO$ 's ability to emphasize the object's shape and to capture its salient features with respect to a shallow color mapping.

[19], [10], [20] and many others. In this setting we used the Nesterov (for Washington and JHUIT-50) and ADAM (for BigBIRD) solvers. As we were only training the last fully connected layer, we were learning only a small handful of parameters; thus, there was a very low risk of overfitting. Upon acceptance of this paper we will release full code for training and evaluation procedures.

#### D. Results

Table I reports the results obtained over the three different databases, for the five different architectures we selected, while Figure 5 and 6 show the class recall on a couple of those experiments. For every architecture, we report the results obtained using ColorJet,  $(DE)^2CO$  learned on a reference database between Washington or JHUIT-50, and  $(DE)^2CO$  learned on the combination of Washington and JHUIT-50. We attempted also to learn  $(DE)^2CO$  from BigBIRD alone, and in combination with one (or both) of the other two databases. Results on BigBIRD only were disappointing, and results with/without adding it to the other

two databases did not change the overall performance. We interpret this result as caused by the relatively small variability of objects in BigBIRD with respect to depth, and for sake of readability we decided to omit them in this work. For every  $(DE)^2CO$  mapping, we report the recognition accuracy only when the colorization was learned from a database different from the one used as testbed for recognition. This is to ensure a fair comparison with ColorJet, as with this protocol both methods use training data from the testbed database only when learning the final classification layer.

We see that, for all architectures and for all reference databases,  $(DE)^2CO$  achieves higher results than ColorJet. The difference goes from a minimum of +0.1%, obtained with CaffeNet on the Washington database, up to +17.4% for VGG16 on JHUIT-50. JHUIT-50 is the testbed database where, regardless of the chosen architecture,  $(DE)^2CO$  achieves the strongest gains in performance compared to ColorJet, followed by BigBIRD. Washington is, for all architectures, the database where ColorJet performs best,

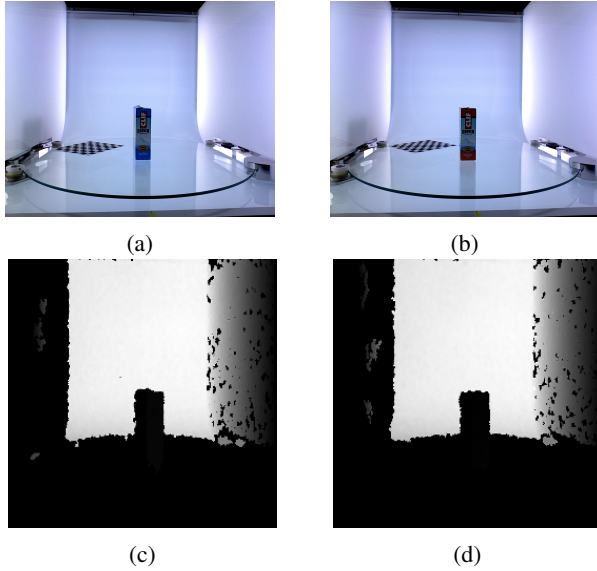


Fig. 4: (a) and (b): pictures from classes *clif crunch peanut butter* and *clif crunch chocolate chips*. (c) and (d): the corresponding normalized depth images. We see that the two objects are indistinguishable in the depth domain (bottom row), being their main difference in color and texture (top row).

with the combination Washington to CaffeNet being the most favorable to the shallow mapping. It is worth noting that indeed ColorJet has been proposed and tested on this specific data+architecture combination, which is by far the most popular combination researched in the literature [19], [10], [20], [5]. It is therefore possible that ColorJet has been inadvertently over-optimized over this combination. On average, it appears that VGG16 is the architecture that leads to the best performances when combined with (DE)<sup>2</sup>CO. This might appear somehow against the computer vision literature, where several results show that architectures like GoogLeNet and ResNet perform better on ImageNet [2], [14]. Still, it should be noted that we are using here all architectures as feature extractors rather than as classifiers. On this type of tasks, both ResNet and GoogLeNet-like networks are known to perform worse than VGG16 [29], hence our results are consistent with what reported in the literature.

We would like to make a final comment about the impact of fine tuning over our (DE)<sup>2</sup>CO results. One might wonder if the significant advantage achieved by our method over ColorJet would still hold after fine-tuning the overall architecture on the testbed databases. Very preliminary investigations using the fine-tuning strategy presented in [10] over two different databases gave conflicting results, with experiments over Washington signaling a substantial equivalence between the two approaches after fine tuning, and experiments over JHUIT-50 still maintaining a strong advantage for (DE)<sup>2</sup>CO. Although these results seems to be in line with the trends reported in Table I, we believe that many more fine tuning

strategies (such as [8]) and all five architectures (and possibly more) should be extensively investigated before taking a stand on this point.

## V. CONCLUSIONS

This paper presented a network for learning deep colorization mappings. Our architecture follows the residual philosophy, learning how to map depth data to RGB images for a given pre-trained convolutional neural network. By using our (DE)<sup>2</sup>CO algorithm, as opposed to shallow depth colorization mapping commonly used in the literature, we obtained a very significant jump in performance over three different benchmark databases, using five different popular deep networks pre trained over ImageNet. The visualization of the obtained colorized images further confirm how our algorithm is able to capture the rich informative content and the different facets of depth data. We plan to make available to the community all the deep depth mappings used in this paper upon acceptance of the paper.

We plan to continue this work in many ways. First, we plan to investigate up to which extent the advantage obtained by learning the colorization mapping affects fine tuning. This in turn will require exploring several fine tuning strategies, as the size of the data and the type of architecture chosen for classification might affect the result and the choice for the best protocol. Second, we showed in [5] that learning a deep architecture from synthetic depth data leads to learning complementary features compared to those learned by a CNN pre-trained over ImageNet. That result should be combined with our current finding, merging the two approaches into a single deep architecture able to take advantage of the different information carried by the two approaches. Future work will focus on these directions.

## REFERENCES

- [1] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems 2*, D. S. Touretzky, Ed. Morgan-Kaufmann, 1990, pp. 396–404.
- [2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.
- [5] F. M. Carlucci, P. Russo, and B. Caputo, "A deep representation for depth images from synthetic data," in *Proc. ICRA*, 2017.
- [6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 647–655.

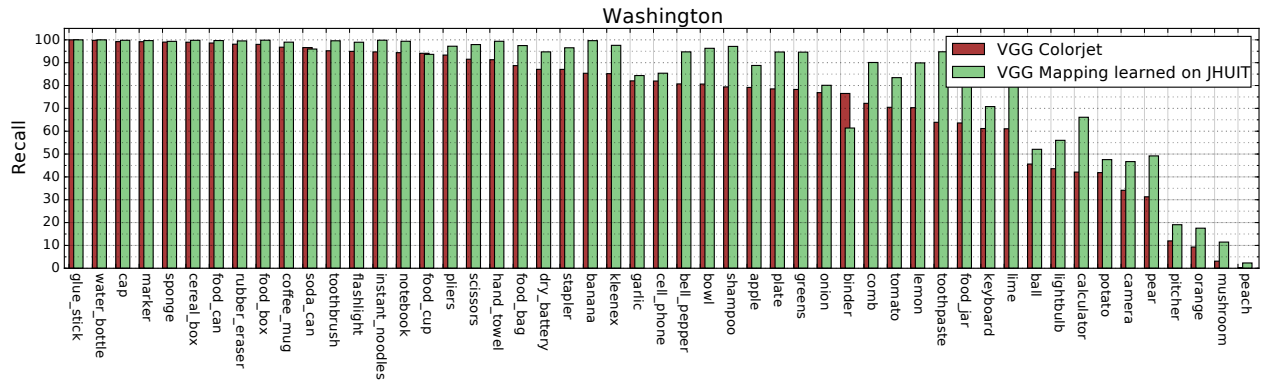


Fig. 5: Per class recall on Washington, using VGG, with  $(DE)^2CO$  learned from JHUIT-50. Recalls per class are sorted in decreasing order, according to the ColorJet performance. With the notable exception of *binder*, we see that  $(DE)^2CO$  outperforms ColorJet on every class.

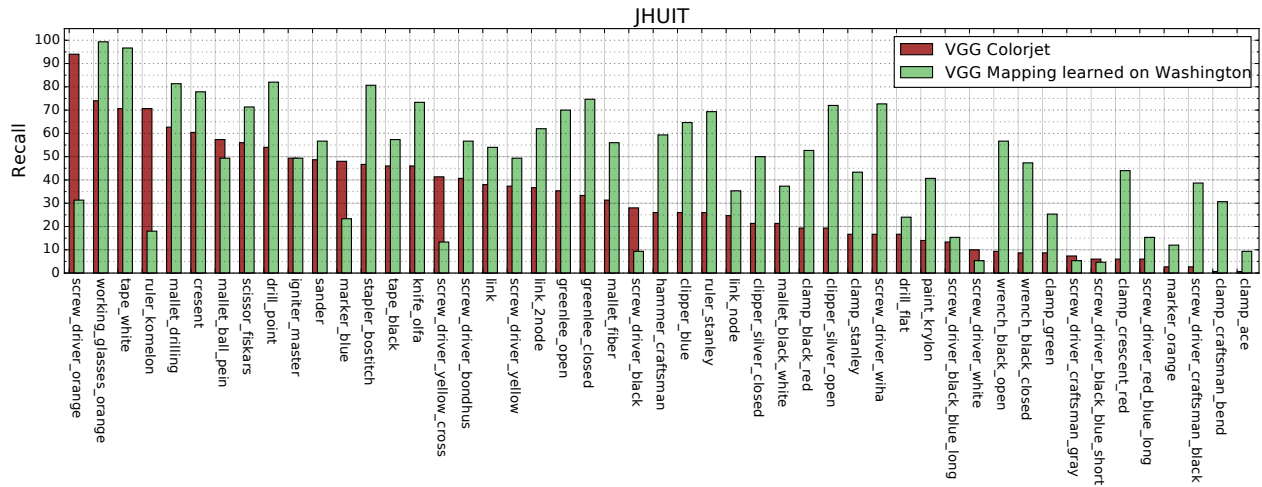


Fig. 6: Per class recall on JHUIT-50, using VGG, with  $(DE)^2CO$  learned from Washington. Recalls per class are sorted in decreasing order, according to the ColorJet performance. In this setting we see that  $(DE)^2CO$ , while generally performing better, seems to focus on different perceptual properties and is thus, compared with the baseline, better at some classes rather than others.

- [8] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 3320–3328. [Online]. Available: <http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks>
- [9] M. Schwarz, H. Schulz, and S. Behnke, “Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 1329–1335.
- [10] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, “Multimodal deep learning for robust rgb-d object recognition,” in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 681–687.
- [11] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification,” *ACM Transac-*

Method:	Washington[24]	JHUIT[25]	BigBIRD Reduced[26]
VGG16 on ColorJet	75.2	35.3	19.9
VGG16 mapping learned on Washington	—	52.7	22.8
VGG16 mapping learned on JHUIT	84.3	—	23.7
VGG16 mapping learned on Washington + JHUIT	—	—	24.0
CaffeNet on ColorJet	78.8	45.0	22.7
CaffeNet mapping learned on Washington	—	53.1	28.6
CaffeNet mapping learned on JHUIT	78.9	—	25.2
CaffeNet mapping learned on Washington + JHUIT	—	—	29.3
GoogLeNet on ColorJet	73.5	40.0	21.8
GoogLeNet mapping learned on Washington	—	51.9	25.2
GoogLeNet mapping learned on JHUIT	76.6	—	24.4
GoogLeNet mapping learned on Washington + JHUIT	—	—	28.6
ResNet50 on ColorJet	75.1	33.5	18.7
ResNet50 mapping learned on Washington	—	45.5	23.9
ResNet50 mapping learned on JHUIT	76.4	—	24.7
ResNet50 mapping learned on Washington + JHUIT	—	—	23.7
SqueezeNet1.1 on colorjet	74.5	37.4	15.9
SqueezeNet1.1 mapping learned on Washington	—	45.9	21.0
SqueezeNet1.1 mapping learned on JHUIT	78.8	—	21.5
SqueezeNet1.1 mapping learned on Washington + JHUIT	—	—	21.7

TABLE I: Object classification experiments in the depth domain, comparing (DE)<sup>2</sup>COand ColorJet, using five networks pre trained on ImageNet as feature extractors.

- tions on Graphics (Proc. of SIGGRAPH 2016), vol. 35, no. 4, 2016.
- [12] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision (ECCV)*, 2016.
- [13] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 415–423.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.
- [15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR 2015*, vol. abs/1409.1556, 2015.
- [16] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," *arXiv:1602.07360*, 2016.
- [17] S. Gupta, R. B. Girshick, P. A. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*, 2014, pp. 345–360.
- [18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [19] M. Schwarz, H. Schulz, and S. Behnke, "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1329–1335.
- [20] H. F. Zaki, F. Shafait, and A. Mian, "Convolutional hypercube pyramid for accurate rgb-d object category and instance recognition," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1685–1692.
- [21] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for RGB-D based object recognition," in *Experimental Robotics - The 13th International Symposium on Experimental Robotics, ISER 2012, June 18-21, 2012, Québec City, Canada, 2012*, pp. 387–402.
- [22] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," *ECCV*, 2016.
- [23] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," *arXiv preprint arXiv:1612.05424*, 2016.
- [24] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1817–1824.
- [25] C. Li, A. Reiter, and G. D. Hager, "Beyond spatial pooling: Fine-grained representation learning in multiple domains," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4913–4922.
- [26] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel, "Bigbird: A large-scale 3d database of object instances," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 509–516.
- [27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 1106–1114.
- [28] Y. Nesterov, "A method of solving a convex programming problem with convergence rate  $O(1/k^2)$ ," in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1983, pp. 372–376.
- [29] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1790–1802, 2016.